Google - tiedonhakijan paras ystävä?

AGRICOLA

Google on muutamassa vuodessa noussut maailman ylivoimaisesti käytetyimmäksi hakukoneeksi ja sen nimestä on tullut jo lähes verkkotiedonhaun symboli. Mihin Googlen huima suosio oikein perustuu? Mitä se merkitsee tiedonhakijan kannalta? Voiko Google säilyttää asemansa myös tulevaisuudessa?

Miksi hakukoneita tarvitaan?

Tim Berners-Leen 1990-luvun alussa kehittämän World Wide Webin perusideoihin kuului se, ettei sillä ollut mitään keskitettyä hallintoa tai keskuspistettä. WWW-palvelimet sijaitsivat enemmän tai vähemmän sattumanvaraisesti eri puolilla maailmanlaajuista Internet-verkkoa, kuka tahansa saattoi julkaista palvelimilla millaisia

dokumentteja tahansa, ja dokumenttien väliset linkitykset muodostivat sekavan verkoston, jolla ei ollut mitään ylhäältä päin määrättyä järjestystä. Jonkin tietyn tiedontarpeen kannalta relevanttien dokumenttien löytäminen oli hyvin vaikeaa ilman etukäteistietoa niiden verkko-osoitteista. Lyhyesti sanottuna World Wide Web oli kasvunsa myötä kehittymässä kohti täydellisen anarkistista kaaosta.

Aivan WWW:n alkuvaiheista lähtien erilaiset linkkilistat olivat yksi keskeinen tapa järjestää verkkoresursseja, ja verkon kasvaessa näiden pohjalta kehittyi vähitellen yhä laajempia hakemistoja ja portaaleita. Esim. Yahoo aloitti toimintansa opiskelijapoikien harrastuksena vuonna 1994, ja siitä tuli nopeasti yksi tunnetuimmista verkkopalveluista. Toinen ja ajan myötä merkittävämpi ilmiö verkkotiedon löydettävyyden kannalta olivat verkkodokumentteja haravoivat hakurobotit ja niiden keräämää aineistoa indeksoivat hakupalvelut. Tämäkään ajatus ei ollut uusi, sillä jo ennen Webin läpimurtoa oli kehitetty mm. ftp-palvelimien sisältämän aineiston ja gopher-resurssien hakuun kykeneviä robotteja. World Wide Webin laajentuessa hakukoneet saivat kuitenkin aivan uudenlaisen merkityksen, sillä verkon nopea kasvu toi mukaan myös monenlaisia taloudellisia intressejä, jotka samalla kertaa loivat mahdollisuuksia aiempaa kunnianhimoisempien teknisten järjestelmien rakentamiseen ja toisaalta toivat hakupalvelujen kehitykselle uusia kaupallisia reunaehtoja.(1)

Googlen nopea nousu

Googlen kiri verkon suosituimmaksi hakukoneeksi tapahtui monella tapaa yllättäen. Google ei näet ollut ollenkaan Webin ensimmäinen hakukone, vaan päinvastoin se aloitti toimintansa verrattain myöhään, siinä vaiheessa kun markkinat näyttivät jo jakautuneen

useiden vuosina 1994-1996 toimintansa aloittaneiden hakupalveluiden kesken. Esim. amerikkalaisen tietokonevalmistaja Digital Equipment Corporationin perustama, loppuvuodesta 1995 toimintansa aloittanut AltaVista oli jo ennättänyt vakiinnuttaa itsensä lähes Googlen kaltaiseksi verkkotiedonhaun synonyymiksi, ja myös Lycos, Excite, HotBot ja Infoseek olivat kaikki varteenotettavia hakupalveluita. Google onnistui siis muutamassa vuodessa kiilaamaan näiden aiemmin aloittaneiden kilpailijoidensa ohi. Vielä hämmentävämpää on se, että Google pystyi saavuttamaan tiedonhakijoiden suosion maailmanlaajuisesti ilman mainittavaa markkinointia. Yahoon ja muiden suurten yritysten kanssa tehdyt sopimukset toivat sille toki runsaasti näkyvyyttä, mutta muuten Googlen suosion kasvu perustui suurelta osin sen tarjoaman hakupalvelun laatuun.

Googlen perustivat vuonna 1998 Stanfordin yliopiston tietojenkäsittelytieteen jatko-opiskelijat Larry Page ja Sergey Brin. Page ja Brin esittelivät kehittelemäänsä hakutulosten indeksointialgoritmiä "The Anatomy of a Large-Scale Hypertextual Web Search Engine" -nimisessä paperissa keväällä 1998,(2) ja saman vuoden syyskuuhun mennessä algoritmi oli patentoitu nimellä *PageRank*. Tässä vaiheessa Googlen hakukone oli jo toiminnassa, ja vaikka se ei vielä sisältänyt muita hakuvaihtoehtoja kuin simppelin yksinkertaisen sanahaun (monipuolisempia hakumahdollisuuksia tarjoava *Advanced Search* -sivu ilmaantui Googlen käyttöliittymään vasta pari vuotta myöhemmin), sana uudesta innovatiivisesta hakupalvelusta levisi nopeasti viidakkorummun välityksellä.(3)

Algoritmiakin tärkeämpää saattoi kuitenkin olla keskittyminen olennaiseen: siinä missä muut hakupalvelut rakensivat itsestään kilvan kaikenkattavia portaalipalveluita, joissa haku oli vain yksi toiminto muiden joukossa, Google pyrki johdonmukaisesti profiloitumaan nimenomaan hakupalveluna. Lisäksi Googlen selkeänä päämääränä oli olla verkon verkon paras ja kattavin hakupalvelu. Siinä missä esim. AltaVistan indeksi alkoi jäädä vuoden 1998 tienoilla yhä pahemmin ajastaan jälkeen (osittain siksi, että AltaVista odotti www-palvelujen ylläpitäjien *maksavan* uusien sivujen lisäämisestä hakupalveluunsa), Google panosti aktiiviseen, palvelujen pääsivuja syvemmälle ulottuvaan haravointiin. Onkin

mielenkiintoinen kysymys missä määrin Googlen saavuttama asema johtuu sen teknisistä innovaatioista, ja missä määrin sen noudattamasta selkeästä, pitkäjänteisestä politiikasta ja politiikan tuloksena syntyneestä hyvästä imagosta.

Yksi tärkeä selitys Googlen hyvälle imagolle on se, ettei se monista kilpailijoistaan poiketen ole ryhtynyt myymään hakutulostensa kärkipaikkoja eniten tarjoaville. Vaikka myös Google myy palvelunsa sivuille tiettyihin hakutermeihin liittyviä ja tietyille kohderyhmille kohdistettuja mainoksia, maksettu materiaali on selkeästi erotettu varsinaisista hakutuloksista, ja se sisältää pelkästään tekstiä, ei kuvia. Tämän vuoksi Googlen käyttöliittymä on pysynyt selkeänä ja helposti hahmotettavana, etenkin jos sitä vertaa monien kilpailevien hakupalveluiden tarjoamiin ylitsepursuaviin portaalisivustoihin.

Google dance ja PageRank

Googlen hakurobotit kiertävät haravoimassa verkkosivustoja säännöllisin väliajoin. Muiden vastaavien robottien tapaan ne seuraavat verkkosivuilla olevia linkkejä ja etenevät siten sivulta toiselle. Periaatteessa Google voi siis näin löytää kaikki sellaiset verkkosivut, joihin on ulkopuolisia linkkejä. Käytännössä tilanne ei ole aivan näin aurinkoinen, sillä nykyään huomattava osa verkkosivuista tuotetaan dynaamisesti erilaisilla tietokantapohjaisilla tekniikoilla, ja tällaisissa tapauksissa kaikkien palvelun sisältämien verkkosivujen haravointi ei aina ole mielekästä tai edes mahdollista.

Hakurobottien kokoamat tiedot kootaan ja järjestetään automaattisesti Googlen indeksiin, joka on hajautettu kymmenille tuhansille pienille Linux-palvelimille. Indeksin koosta johtuen tietojen päivittyminen tapahtuu indeksin eri osissa eri aikaan, ja päivitysprosessin aikana Googlen antamat hakutulokset saattavat vaihdella hyvinkin arvaamattomasti. Väitetään myös, että osoitteiden www2.google.com ja www3.google.com kautta löytyvät indeksit poikkeavat Googlen varsinaisesta indeksistä, ja osaltaan ennakoivat tietojen päivittymistä. Google on yleensä päivittänyt koko indeksinsä noin kuukauden välein. Tärkeimpiä ja useimmin päivittyviä sivuja on tosin jo muutaman vuoden ajan haravoitu tiuhempaan tahtiin. Lisäksi Google on perustanut uutispalvelujen ja verkkolehtien reaaliaikaista seurantaa varten erillisen hakupalvelun (Google News).

Googlen päivitysprosessi tunnetaan www-palvelujen ylläpitäjien keskuudessa nimellä *Google dance*. Googlen tanssi on jo muutamassa vuodessa ennättänyt saada verkkomaailmassa lähes samanlaisen myyttis-uskonnollisen merkityksen kuin esim. Niilin tulvat muinaisessa Egyptissä. Sitä sekä odotetaan että pelätään, sillä etenkin kaupallisten sivustojen ylläpitäjille on elämän ja kuoleman kysymys onko heidän palvelunsa tärkeimpien hakutermien tuottamissa tuloksissa ensimmäisellä sivulla tai peräti ensimmäisenä, vai löytyykö se esim. kymmenenneltä sivulta tai onko se tipahtanut kokonaan hakutuloksista.(4) Uusi indeksi voi siis sekä tuoda mukanaan uutta liikennettä ja uusia asiakkaita että viedä vanhatkin pois. Koska Googlen käyttämä algoritmi ja indeksointipolitiikka ovat (muiden hakukoneiden tavoin) liikesalaisuuksia, tuloksissa tapahtuneille muutoksille on usein vaikea löytää varmaa selitystä.(5) Niinpä hakukoneoptimoinnista (*Search Engine Optimization, SEO*) on kasvanut oma kukoistava bisnesalansa/salatieteensä, ja Googlen ja muidenkin hakukoneiden ominaisuuksia analysoidaan esim. sellaisilla www-sivustoilla kuin WebmasterWorld tai Search Engine World.

Googlen alkuperäinen tekninen innovaatio PageRank liittyy hakurobotin keräämän aineiston indeksointiin ja hakutulosten järjestämiseen. Verkosta löytyvien dokumenttien määrä oli jo vuoteen 1998 mennessä kasvanut niin suureksi, että yleisimmillä hakusanoilla osumia kertyi kymmeniä tai jopa satojatuhansia. Niinpä yhä oleellisemmaksi kysymykseksi oli tullut se, miten hakutulokset järjestetään relevanssin mukaan niin, että käyttäjän kannalta kiinnostavimmat dokumentit ovat listauksessa ensimmäisinä. Suurin osa tiedonhakijoista ei näet ollut kiinnostunut selaamaan loputtomalta vaikuttavia hakutuloslistauksia läpi kiinnostavien dokumenttien toivossa, vaan parhaat ja relevanteimmat hakutulokset oli saatava tavalla tai toisella heti tuloslistan alkuun, mielellään jo ensimmäiselle tulossivulle. AltaVista ja muut varhaiset hakukoneet pyrkivät yleensä ratkaisemaan ongelman kirjastotietokannoista tutun boolean-logiikan

avulla. Käytännössä kuitenkin vain pienellä osalla tiedonhakijoista riitti mielenkiintoa boolean-operaattorien käytön opetteluun.

Lisäongelmia aiheutti myös se, että hakukoneet yrittivät hyödyntää keräämänsä materiaalin indeksoinnissa www-sivujen otsikkotietoihin sisältyviä metadata-kenttiä. Ikävä kyllä näitä kenttiä käytettiin vähitellen yhä yleisemmin pikemminkin hakukoneiden hämäämiseen kuin tiedonhaun helpottamiseen. Sivujensa kävijämäärien kasvattamiseksi näet monet www-sivujen ylläpitäjät ryhtyivät lisäämään sivujensa metadata-kenttiin kaikkein suosituimpia hakutermejä (tyypillisinä esimerkkeinä mm. *seksi* ja *porno*), riippumatta siitä oliko sivulla mitään ko. hakusanoihin liittyvää materiaalia. Käytännössä tämä johti ennen pitkää siihen, että sivuntekijöiden itsensä laatima metadata muuttui dokumenttien indeksoinnin kannalta hyödyttömäksi.

Googlen käyttöön ottamassa indeksointialgoritmissa oli olennaista se, että Page ja Brin päättivät olla noteeraamatta sivuntekijöiden laatimaa metadataa, ja sen sijaan Google keskittyi analysoimaan toisaalta suoraan sivujen tekstisisältöä ja toisaalta sivujen välisten linkkien muodostamia suhteita. Googlen käyttämän logiikan mukaan sivu on sitä merkittävämpi mitä enemmän siihen on linkkejä, ja nämä linkit ovat sitä arvokkaampia mitä merkittävämmiltä sivuilta ne tulevat. Sivujen sijoittuminen hakutuloksissa määräytyy toisaalta PageRankin määrittelemän sivun yleisen arvon ja toisaalta sen hakutermiin liittyvän relevanssin perusteella. Myös sivuun viittaavien linkkien tekstillä on merkitystä sen indeksoitumiselle - se, että jokin sivu esiintyy hakutuloksissa vaikkei hakutermiä mainita sivulla lainkaan voi selittyä tätä kautta.(6)

Käytännössä kunkin verkkosivun PageRank esitetään yleensä kokonaislukuna nollan ja kymmenen väliltä.(7) Vain muutamien verkon tärkeimpien sivujen (mm. Googlen oma kotisivu) PageRank on kymmenen, ja suurin osa muista sivuista saa huomattavasti alemman arvon. Google tiettävästi suosii yliopistoja yms. julkisia palveluita indeksoinnissaan, ja niinpä monien yliopiston palvelimilla sijaitsevien sivujen PageRank saattaa olla korkeampi kuin vastaavilla muualla sijaitsevilla sivuilla. Esim. Helsingin yliopiston kotisivu on ollut Googlen arvottamana seitsemän arvoinen, yliopiston alasivuilla PageRank laskee vähitellen kuudesta alaspäin. Monet kaupalliset palvelut voivat vain unelmoida tällaisista lukemista.

Googlea on kritisoitu siitä, että sen käyttämä algoritmi ohjaa käytännössä lisää liikennettä jo ennestään suosituille sivuille, ja siten voimistaa entisestään verkkosivujen käytön keskittymistä lähes pelkästään tietyille sivustoille.(8) On tosin tulkinnanvaraista onko tämä yksinomaan huono asia, sillä todennäköisesti Googlen tarjoamat tulokset vastaavat kuitenkin useimpien tiedonhakijoiden toiveita. Käytännössä tämä on voinut johtaa siihen, että uusilla sivuilla voi olla sisältönsä merkittävyydestä huolimatta vaikeuksia sellaisten vanhojen ja vakiintuneiden sivustojen ohittamisessa, joihin on jo olemassa runsaasti linkkejä.(9) Kannattaa kuitenkin huomata, että tällaiset kysymykset tulevat yleensä esiin vain silloin, kun tehdään hakuja jollain sellaisella *kilpaillulla* hakutermillä, joka tuottaa tulokseksi suuren määrän sivustoja. Harvinaisempien hakutermien kohdalla riittää, että sivu on ylipäänsä indeksoitu, ja tässä suhteessa Google kuuluu tehokkaan haravointinsa ansiosta hakukoneiden parhaimmistoon.

Google on pyrkinyt järjestelmällisesti laajentamaan hakuaan myös sellaisiin aineistoihin, jotka ovat aiemmin kuuluneet ns. näkymättömään webiin. Tavallisten HTML-muotoisten sivujen lisäksi Google on ryhtynyt indeksoimaan myös monissa muissa formaateissa tallennettuja aineistoja, esim. PDF- ja Word-tiedostoja. "Tavallisiin" verkkosivuihin kohdistuvan hakupalvelun lisäksi Google on perustanut mm. kuvatiedostoihin ja verkon uutispalveluihin kohdistuvat hakupalvelut sekä liittänyt palveluihinsa alun perin DejaNewsin 90-luvun puolivälistä lähtien keräämän ja Googlen sittemmin ostaman ja myös taannehtivasti laajentaman Usenetin keskusteluryhmien arkiston (Google Groups).(10) Muitakin erikoistuneita hakupalveluita on joko suunnitteilla tai jo testattavana,(11) ja lisäksi Google on satsannut myös esim. kieliteknologiaan.(12)

Uusien aineistojen ja uusien hakupalveluiden lisäksi Google on laajentunut myös perustamalla lukuisia kansallisia ja erikielisiä versioita hakupalveluistaan. Esim. suomalaisille käyttäjille suunnatun www.google.fi:n indeksi poikkeaa jonkin verran www.google.com:in indeksistä, minkä lisäksi suomenkielisen Googlen palveluvalikoimasta puuttuu amerikkalaiseen versioon sisältyvä uutishaku. Räätälöityjen hakupalveluiden lisäksi Googlen

kansallisia versioita selittävät epäilemättä myös liiketaloudelliset syyt: kansalliset versiot tarjoavat parempia mahdollisuuksia hakutulosten yhteyteen liitettävien tekstimainosten kohdistamiseen.

Haussa hyvä bisnesidea...

Google on muutamassa vuodessa kyennyt luomaan itsestään maailmanlaajuisesti tunnetun *brandin*.(13) Tästä huolimatta ei ole ollut ollenkaan itsestään selvää miten tällaisen aseman voi muuttaa rahaksi - maailmanlaajuisessa käytössä olevan hakukoneen ylläpitäminen on näet sinällään erinomaisen huono bisnesidea. Jo pelkkä hakupalvelun teknisen infrastruktuurin ylläpito vaatii väistämättä runsaasti rahaa, etenkin jos palvelun suosio on Googlen luokkaa. Niinpä yksi keskeisimmistä Googleen liittyvistä kysymyksistä onkin, millaisen liiketoimintamallin ympärille se tulevaisuutensa rakentaa. Niin kauan kuin hakukoneen käyttö ja siihen listautuminen ovat molemmat ilmaisia, palvelun rahoitus on väistämättä riippuvainen joko palvelun kautta myytävistä mainoksista, sen tarjoamista maksullisista oheispalveluista tai muiden yritysten kanssa tehtävistä yhteistyösopimuksista.(14)

Vaikka Google aloitti toimintansa periamerikkalaiseen tapaan autotallista, se ei ole suinkaan syntynyt pelkän Pyhän Hengen voimalla. Google sai toimintansa alkuvaiheessa huomattavan määrän rahoitusta (kymmeniä miljoonia dollareita) useilta amerikkalaisilta riskisijoittajilta, jotka edelleen omistavat suuren osan yrityksestä. Lisäksi Google onnistui varhaisessa vaiheessa myymään hakutuloksensa mm. Yahoon käyttöön, mikä jo sinällään takasi kohtuullisen kassavirran.(15)

Vaikka Googlen listautumista pörssiin on odotettu jo parin vuoden ajan, se on toistaiseksi pysynyt yksityisesti omistettuna yhtiönä. Listautuminen olisi todennäköisesti tehnyt yhtiön nykyisistä omistajista upporikkaita ja tuonut myös yhtiön käyttöön runsaasti rahaa, mutta Googlen valitsema strategia lienee kuitenkin ollut järkevä toiminnan pitkäjänteisen kehittämisen kannalta. Varoittavia esimerkkejä toisenlaisten valintojen sisältämistä riskeistä löytyy yllin kyllin: Listautumiseen liittyneet ja osakekursseihin kasautuneet huimat kasvu- ja tulosodotukset houkuttelivat näet monet Googlen kilpailijoista lyhytnäköisiin palvelun kaupallistamiseen tähdänneisiin ratkaisuihin. Nämä palvelivat yleensä pikemminkin

osakkeenomistajien oletettua etua kuin palvelun käyttäjiä, ja vahingoittivat siten sekä palveluiden käytettävyyttä että imagoa.(16) Toisaalta on ilmeistä, että Googlekin joka tapauksessa listautuu pörssiin ennen pitkää, sillä on vaikea uskoa, etteivät Googlea rahoittaneet sijoittajat haluaisi jossain vaiheessa realisoida omistuksiaan. Tämä vaihe tulee epäilemättä olemaan kriittinen sekä Googlen tulevan kehityssuunnan että sen imagon kannalta.

Hakukoneiden välinen kilpailutilanne on muuttunut parin viime vuoden aikana, kun Googlen jalkoihin jääneet kilpailijat ovat yksi toisensa jälkeen päätyneet jonkin muun ostajan haltuun. Hakupalveluiden omistus on kasautunut niin, että aiemmin linkkihakemistostaan tunnettu Yahoo omistaa nykyisin sekä Inktomin, Overturen, AltaVistan että norjalaisen FAST:in alun perin kehittämän AllTheWebin. Hakukonemarkkinat näyttävätkin näillä näkymin olevan kolmen kauppa Googlen, Yahoon ja omaa hakukonettaan kehittelevän Microsoftin kesken.

Etenkin Microsoftin odotettavissa oleva tulo mukaan kilpailuun koventaa panoksia, sillä yhtiöllä on käytössään lähes rajattomat resurssit ja myös merkittäviä kilpailuetuja johtavana käyttöjärjestelmien, toimisto-ohjelmien ja www-selainten valmistajana. Kuten tunnettua, yhtiö ei ole arastellut käyttää johtavaa markkina-asemaansa hyväkseen, ja Bill Gates ennätti jo tammikuussa 2004 julistamaan, että Microsoft tulee kirimään Googlen etumatkan kiinni. Monet ennustelevat jo Microsoftin ja Googlen välille "hakukoneiden sotaa", käyttäen vertailukohtanaan Microsoftin Internet Explorer -selaimen ja Netscapen 90-luvun loppupuolella käymää kamppailua, joka päättyi Microsoftin voittoon.(17) Microsoftin valitsemassa strategiassa herättää tosin ihmetystä se, ettei yhtiö ostanut yhtään aiemmista hakukoneyrityksistä vaan antoi niiden ajautua muiden käsiin. Sen sijaan Microsoft on testannut omaa hakurobottiaan, joka lienee edelleen melko keskeneräinen.

Yahoolla puolestaan on ostoksiensa kautta hallussaan runsaasti perinteikkäitä hakupalveluita, joista etenkin AllTheWebin hakuteknologia ja indeksi on vetänyt vertoja jopa Googlelle. Toisaalta osa Yahoon omistamista hakupalveluista (etenkin nimenomaan maksettuihin listauksiin erikoistunut Overture) on profiloitunut myymällä hakutulostensa kärkipaikkoja, mikä vähentää niiden tarjoamien hakumahdollisuuksien uskottavuutta

tiedonhakijoiden näkökulmasta. Vielä toistaiseksi Yahoo hyödyntää omassa palvelussaan Googlen indeksiä, mutta se on jo ilmoittanut siirtyvänsä käyttämään omaa hakuteknologiaansa alkuvuoden 2004 aikana.(18)

Google lähtee kilpailuun joka tapauksessa johtoasemasta. Epäilemättä myös Googlelle löytyisi yllin kyllin kiinnostuneita ostajia, mutta ainakin tämänhetkisten tietojen mukaan yhtiö "ei ole myytävänä". Pörssilistautuminen saattaa toki muuttaa asetelmia tässäkin suhteessa.

Isoveli valvoo?

Useimpien arvioiden mukaan jo yli puolet nettihakukoneiden käyttäjistä käyttää nimenomaan Googlea, toisinaan Googlen markkinaosuuden arvioidaan olevan jo lähempänä kolmea neljäsosaa. Google on noussut verkkotiedon haussa ainakin tilapäisesti lähes samanlaiseen ylivoimaiseen valta-asemaan kuin mikä Microsoftilla on tietokoneiden käyttöjärjestelmissä ja toimisto-ohjelmissa. Niinpä ei olekaan ihme, että Googlea ja sen toimintatapoja kohtaan on alkanut esiintyä myös jonkin verran kritiikkiä, vaikkei kukaan kiistäkään sen käyttökelpoisuutta hakukoneena. Mitään Microsoftin herättämiä tunteita vastaavaa vastarintaliikettä on tuskin näköpiirissä, mutta ylivoimainen markkinajohtaja herättää lähes väistämättä myös ärtymystä, ja hyvätkin aikomukset saatetaan helposti tulkita ylimielisyydeksi.

Esimerkiksi Google Watch -nimiseltä verkkosivustolta löytyy kokoelma Googlea kritisoivia tekstejä, joissa kyseenalaistetaan palvelun luomaa positiivista imagoa ("Google is not your friend"). Sivuilla esitetään myös useita aivan aiheellisia kysymyksiä siitä, miten Google kerää tietoja tiedonhakijoista ja heidän tekemistään tiedonhauista ja mihin se näitä tietoja käyttää. Sivujen mukaan Googlen käyttäjilleen tarjoamassa yksityisyyden suojassa on paljon

toivomisen varaa. Daniel Brandtin kirjoittamassa artikkelissa kiinnitetään huomiota mm. siihen, että Google asentaa käyttäjän kovalevylle pysyvän keksitiedoston (*cookie*), jonka avulla tämän tekemät tiedonhaut voidaan ainakin periaatteessa yksilöidä ja yhdistää yhtenäiseksi profiiliksi.(19) Vaikka Google ei tällä hetkellä käyttäisi keräämäänsä aineistoa väärin, onko mitään takeita siitä, että näin olisi myös tulevaisuudessa?

Tällaiset "Isoveli valvoo" -tyyppiset skenaariot ovat luonnollisesti saaneet erityistä lisäpainoa syyskuun 11. päivän 2001 tapahtumien ja sen jälkeen alkaneen "terrorisminvastaisen sodan" myötä. Voisi kuvitella, että Googlen kaltainen maailmanlaajuisessa käytössä oleva hakupalvelu olisi monin tavoin kiinnostava kohde potentiaalisia epäiltyjä etsiville tiedustelupalveluille.(20) Jos käyttäjien tekemät tiedonhaut ovat yksilöitävissä, tämä tarjoaa monenlaisia mahdollisuuksia erilaisten henkilörekisterien rakentamiseen ja epäilyttävien tiedonhakujen ja -hakijoiden tarkkailuun. Tällaisia visioita on itse asiassa tullut mieleen myös monista Googlea kuvaavista positiivissävyisistä artikkeleista, joissa kerrotaan yhtiön toimitilojen ala-aulassa olevasta taulusta, jolle heijastetaan reaaliajassa käyttäjien tekemiä hakuja. Onko Google siis yksi potentiaalinen askel kohti historioitsija Paul N. Edwardsin kuvaamaa "suljettua maailmaa",(21) maailmanlaajuista taistelukenttää, jossa palvelun käyttäjien tekemät tiedonhaut voivat päätyä osaksi turvallisuuspalveluiden keräämiä rekistereitä?
Turvallisuuspalveluiden ohella toinen ilmeinen tiedonhauista kiinnostunut taho ovat erilaiset kauppaketjut ja markkinointiyritykset, jotka jo nyt keräävät rutiininomaisesti monenlaista informaatiota ihmisten mielenkiinnon kohteista ja kulutustottumuksista.

Hieman arkisemmalla tasolla yksityisyyden suojaan liittyy myös se, että Googlen kaltaisella hakukoneella kuka tahansa voi helposti koota ja yhdistellä verkosta runsaasti sellaista yksittäistä henkilöä koskevaa materiaalia, jota ei välttämättä ole tarkoitettu yleiseen käyttöön. Tämä ei tietenkään ole millään lailla vain Googlen ominaisuus, vaan jokseenkin väistämätöntä nykyisessä verkottuvassa maailmassa. Joka tapauksessa tämä johtaa siihen, että ihmisten täytyy kiinnittää entistä enemmän huomiota siihen, millaista tietoa he itsestään verkossa antavat.

Toinen merkittävä ongelma on se, miten objektiivisen kuvan Googlen indeksi antaa tarjolla olevasta tiedosta. Vaikka se perustuu näennäisesti automaattiseen aineiston keruuseen ja matemaattiseen algoritmiin, taustalla on kuitenkin väistämättä monenlaisia eri tavoin perusteltuja päätöksiä siitä millaista materiaalia indeksiin halutaan, ja miten näkyvän aseman se saa hakutuloksissa.(22) Googlen kohdalla hakutulosten objektiivisuus kyseenalaistettiin näkyvimmin pari vuotta sitten, jolloin se taipui Skientologia-kirkon lakimiesten painostuksesta tilapäisesti poistamaan indeksistään skientologien pyhiä kirjoituksia ilman lupaa julkaisseita sivuja.(23) Tapaus herätti voimakkaan vastareaktion, josta Google näytti ainakin sillä kertaa ottavan opikseen.

Tällaiset hakutulosten objektiivisuuteen ja yksityisyyden suojaan liittyvät ongelmat eivät tietysti liity yksin Googleen, vaan samoja argumentteja voi kohdistaa myös muihin vastaaviin hakupalveluihin. Googlen asema maailman suosituimpana hakukoneena antaa sille kuitenkin runsaasti valtaa, johon liittyviä ongelmia korostaa se, etteivät yhtiön tekemät hakualgoritmiaan ja indeksiään koskevat päätökset ole yleensä millään lailla julkisia tai läpinäkyviä.

Google ja kirjastot?

Joskus vuosi-pari
sitten olin paikalla
tilaisuudessa, jossa
eräässä
puheenvuorossa
esitettiin Google ja
kirjastot toistensa
kilpailijoina, jopa
siinä valossa, että kirjastot ovat häviämässä kamppailun Googlen tarjoamia hakumahdollisuuksia vastaan. Jäin miettimään sitä, ovatko Googlen kaltaiset hakupalvelut ja kirjastot todella "vihollisia", vai

ovatko ne potentiaalisia liittolaisia? Kilpailevatko hakukoneet ja kirjastotietokannat asiakkaiden sieluista, vai onko pikemminkin niin, että ne täydentävät toisiaan ja palvelevat kumpikin omalla tavallaan tiedonhakijoiden tarpeita?

Kirjastomaailman ja Googlen kaltaisia hakupalveluita kehittävien tietojenkäsittelytieteilijöiden valitsemien keinojen välillä on joka tapauksessa ollut sekä teknisiä että paradigmaattisia eroja. Siinä missä kirjastoihmisten ovat uskoneet mahdollisimman korkeatasoisten kokoelmien mahdollisimman laadukkaaseen luettelointiin, tietojenkäsittelytieteen piirissä pyritään ennemminkin analysoimaan automaattisesti laajoja tekstimassoja. Käytännössä tämä on tarkoittanut sitä, että kirjastoissa on panostettu ihmisvoimin tuotettuun metadataan ja sisällönkuvailuun, kun taas Googlen kaltaiset palvelut pystyvät indeksoimaan automaattisesti jopa koko World Wide Webin laajuisia aineistoja, joiden luettelointi olisi käytännössä mahdotonta millään järjellisissä rajoissa olevilla resursseilla.

Käytännössä näiden lähestymistapojen väliset raja-aidat eivät kuitenkaan ole niin selkeät kuin edellä esitellystä voisi kuvitella, ja voi jopa väittää, että kumpikin leiri on vähitellen omaksumassa toisen ajatuksia. Esim. Kimmo Tuominen on kiinnittänyt huomiota siihen, että Google käyttää jo nyt hakutulostensa tuottamiseen ihmisvoimin tuotettua metadataa, sillä se hyödyntää indeksoinnissaan Open Directory Projectin tuottamaa hakemistoa.(24) Toisaalta myös kirjastoissa on erilaisten digitaalisten aineistojen yleistyessä alettu yhä enemmän kiinnostua dokumenttien automaattisten indeksoinnin tarjoamista mahdollisuuksista.

Oleellisin kirjastojen ja Googlen kaltaisten hakukoneiden ero on edelleen siinä, että ne tarjoavat tiedonhakijan käyttöön varsin erilaisia aineistoja, jotka menevät vain osittain päällekkäin. Kuten esim. Tero Karasjärvi totesi Agricolan tietosanomien edellisessä numerossa, pelkästään Googlen tarjoamien verkkolähteiden varassa historiantutkija ei useimmiten pääsisi aiheensa kanssa kovinkaan syvällisiin tuloksiin.(25) Vaikka Google-haku saattaa tuntua petollisen helpolta ja kattavalta, yleensä kirjaston hyllystä ja kirjastotietokannoista löytyy edelleen moninkertainen määrä tutkimusaiheen kannalta relevanttia tietoa.

Perinteisten painettujen kirjojen ja lehtien ohella etenkin yliopistojen kirjastot tarjoavat käyttäjilleen myös laajan valikoiman erilaisia maksullisia tietokantoja ja aineistokokoelmia, joiden sisältämä materiaali on kaupallisista syistä ainakin toistaiseksi kuulunut verkon yleisten hakukoneiden näkökulmasta niiden tavoittamattomissa olevaan näkymättömään webiin. Kirjastot ovatkin rakentamassa aineistojen hallintaa ja hakua varten omia järjestelmiään. Hyvä esimerkki tällaisesta hankkeesta on esim. Suomessa tämän vuoden kuluessa käyttöön tuleva kansallinen Nelli-portaali, jonka tavoitteena on selkiyttää etenkin FinELibin hankkimien laajojen elektronisten aineistojen hakumahdollisuuksia.(26)

Toisaalta myös Google on alkanut viime aikoina osoittaa kasvavaa kiinnostusta perinteisten kirjastoaineistojen indeksointia kohtaan. Google on jo käynnistänyt pilottihankkeena yhteistyön useiden merkittävien amerikkalaisten kirjankustantajien kanssa: tavoitteena on mahdollistaa kirjojen sisältönäytteisiin, arvioihin ja bibliografisiin tietoihin kohdistuvat haut Googlen hakukoneella.(27) Tuoreimpien huhujen mukaan Googlella on myös yhteistyöprojekti Stanfordin yliopiston kirjaston kanssa: tämän "Projekti valtamerenä" (*Project Ocean*) tunnetun hankkeen huikeana tavoitteena on tiettävästi digitoida kaikki Stanfordin kokoelmiin kuuluvat tekijänoikeudesta vapaat (ennen vuotta 1923 painetut) kirjat ja saattaa ne haettaviksi Googlen kautta.(28)

Jyrki Ilva
Kirjoittaja työskentelee suunnittelijana
Helsingin yliopiston kirjaston tietokantapalveluissa

Lähdeviitteet

(1) Hakukoneiden varhaishistoriasta ks. esim. Sonnenreich 1997 ja Sherman & Price 2002, s. 3-16. Myöhemmästä kehityksestä ks. Olsen & Hu 2003.

(2) Brin & Page 1998.

(3) Ks. myös yhtiön verkkosivuilta löytyvä lyhyt historiikki (http://www.google.com/corporate/history.html).

(4) Google saattaa rankaista sellaisia sivustoja, jotka käyttävät

sijoituksensa parantamiseen epärehellisinä pidettyjä keinoja. Tyypillisiä hakukoneiden hämäämiseen käytettyjä keinoja ovat esim. *cloaking* (palvelin syöttää hakurobotille toisenlaisen sivun kuin tavallisille käyttäjille) tai erilaiset *linkkifarmit* (palvelun ylläpitäjä luo eri osoitteissa sijaitsevien toisiinsa viittaavien sivujen verkoston, jolla yritetään nostaa sivun PageRankia). Google on muiden hakukoneiden tavoin ajoittain muuttanut algoritmiaan karsiakseen erilaisia huijausmahdollisuuksia.

(5) Ks. esim. Olsen 2002.

(6) Toinen todennäköinen selitys on se, että sivu on ennättänyt jo muuttua haravoinnin ja indeksoinnin jälkeen. Tällöin alkuperäinen indeksoitu dokumentti voi kuitenkin yhä löytyä klikkaamalla hakutuloksissa Googlen ylläpitämään välimuistiin (*cache*) johtavaa linkkiä.

(7) Sivujen PageRankin voi selvittää asentamalla koneelleen Google Toolbar -työkalun. Kannattaa kuitenkin huomata, että Google Toolbar lähettää tiedon käyttäjän selaamista sivuista eteenpäin Googlelle, eli se on potentiaalinen tietoturvariski. PageRankista yleisemmin ks. Brin & Page 1998, Brandt 2002, Horrell 2001 ja Craven sa.

(8) Brandt 2002 ja Olsen 2002.

(9) Toisaalta Googlen algoritmillä on ollut aivan päinvastaisia ongelmia viime aikoina yleistyneiden nettipäiväkirjojen eli weblokien (*weblog*, *blog*) kanssa, sillä niiden keskinäiset viittaukset toisiinsa ovat saattaneet hyvin nopeasti nostaa hakutulosten kärkeen hyvinkin sattumanvaraisia sivuja. Ks. esim. Orlowski 2003.

(10) Ilva 2002.

(11) Googlen muista palveluista esim. kaupallisiin ostossivustoihin keskittyvä *Froogle*-hakukone (http://froogle.google.com/) on edelleen betatesti-vaiheessa. Hakupalveluiden lisäksi Google on laajentanut toimintaansa myös ostamalla weblokeja varten luodun *Blogger*-palvelun (http://www.blogger.com). Tuorein esimerkki Googlen rönsyilystä uusille alueille on tammikuussa 2004 toimintansa aloittanut verkkoyhteisö *orkut.com*.

(12) Googlen automaattinen käännöspalvelu (http://www.google.com

/language_tools => Translate) mahdollistaa yllättävän sujuvan alkukielestä käännettyjen sivujen selailun. Palvelun kielivalikoima on tosin edelleen melko rajoittunut, ja kuten tavallista, siitä puuttuu mm. suomi.

(13) Esim. konsulttiyhtiö Interbrandin tekemässä, laajalti uutisoidussa haastattelututkimuksessa Google äänestettiin äskettäin jo toisen kerran peräkkäin maailman arvostetuimmaksi brandiksi. Reuters 2004.

(14) Tämän alaluvun sisältämät tiedot Googlen liiketoimintamalliin liittyvistä suunnitelmista perustuvat suurelta osin artikkeleihin Hansell 2002 ja Knowledge@Wharton 2003.

(15) Tiettävästi Google on pari viime vuotta ollut erittäin voitollinen yhtiö: miljardin dollarin vuosittaisella liikevaihdolla voittoa on kertynyt 350 miljoonaa dollaria. Listautumisen jälkeiseen tulevaisuuteen kohdistuviin odotuksiin verrattuna nämä ovat tosin vielä melko vähäisiä summia.

(16) Sittemmin Googlen haltuun päätynyt Usenetin uutisryhmien viestejä arkistoinut DejaNews on hyvä esimerkki siitä miten kauas lyhytnäköiset palvelun kaupallistamiseen tähtäävät pyrkimykset saattoivat viedä alkuperäisestä toimintaideasta. Deja Newsin vaiheista ks. tarkemmin Ilva 2002.

(17) Markoff 2004. Tiettävästi Microsoft on jo yrittänyt aggressiivisesti värvätä Googlen henkilökuntaa omaan palvelukseensa ja pyrkinyt muutenkin kartoittamaan Googlen heikkoja kohtia.

(18) Hansen & Hu 2004.

(19) Brandt 2003. Kohtuuden nimissä on syytä korostaa, että Googlen oma tämänhetkinen [helmikuu 2004] Privacy Policy (ks. http://www.google.com/privacy.html) antaa asiasta varsin toisenlaisen kuvan. Ks. myös Manjoo 2002 ja Olsen 2002.

(20) Toki kannattaa muistaa, että esim. USA:n tiedustelupalveluilla oli omat verkkoliikenteen tarkkailuun tarkoitetut järjestelmänsä jo kauan ennen Googlea.

(21) Edwards 1997.

(22) Lazuly 2003. Hakutulosten objektiivisuuteen kohdistuvat riskit ovat luonnollisesti korostuneet sitä mukaa kun itsenäisten hakupalveluiden määrä on vähentynyt.

(23) Ks. esim. Gallagher 2002.

(24) Tuominen 2003. Open Directory Projectista ks. http://dmoz.org.

(25) Karasjärvi 2003.

(26) Nelli-portaalista ks. Rouvari & Ryhänen 2003.

(27) Olsen 2003. Palvelun betaversiota voi jo testata rajoittamalla haun palvelimeen print.google.com ja valitsemalla haettavaksi jonkin sopivan hakusanan (esim. "Melville"). Olsen kiinnittää huomiota siihen, että tämä palvelu kilpailee Amazon.comin tarjoaman vastaavan hakumahdollisuuden kanssa.

(28) Markoff 2004. Tässä vaiheessa on tosin epäselvää, tulevatko kirjat kokonaisuudessaan ilmaiseksi luettavaksi verkon kautta, vai onko tavoitteena jonkinlaisen maksullisen palvelun luominen. Jälkimmäinen vaihtoehto lienee digitoinnin ja palvelun ylläpidon vaatimat resurssit huomioiden todennäköisempi.

Kirjallisuusluettelo:

Brandt 2002
Daniel Brandt: PageRank. Google's Original Sin. August 2002.
URL: http://www.google-watch.org/outdated/pagerank.html

Brandt 2003
[Daniel Brandt]: And Then There Were Four: Why We Target Google. [2003]
URL: http://www.google-watch.org/bigbro.html

Brin & Page 1998
Sergey Brin & Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Papers Presented at the 7th International World Wide Web Conference, Brisbane, Australia, 14-18 April, 1998.

URL: http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

Craven sa
Phil Craven: Google's PageRank Explained and how to make most of it. [sine anno]
URL: http://webworkshop.net/pagerank.html

Edwards 1997
Paul N. Edwards: The Closed World. Computers and the Politics of Discourse in Cold War America. Cambridge, MA 1997 (1996).

Gallagher 2002
David Gallagher: Google Runs Into Copyright Dispute. New York Times. April 22, 2002.
URL: http://www.searchutilities.com/news/archive/53/1549.html

Hansell 2002
Hansell, Saul: Google's Toughest Search Is for a Business Model. New York Times. April 8, 2002.
URL: http://www.taipeitimes.com/News/bizfocus/archives/2002/04/10/131343

Hansen & Hu 2004
Evan Hansen & Jim Hu: Yahoo, Google primed for search war. CNET News.com, Jan 14, 2004.
URL: http://news.com.com/2100-1024-5141328.html

Horrell 2001
Mark Horrell: Pagerank Calculator. November 28, 2001.
URL: http://www.markhorrell.com/seo/pagerank.html

Ilva 2002
Jyrki Ilva: Valoa nettikulttuurin alkuhämärään. Agricolan tietosanomat 4/2001. (2002)
URL: http://agricola.utu.fi/tietosanomat/numero4-01/googlenarkisto.html

Karasjärvi 2003
Tero Karasjärvi: Internetissähän se kaikki tieto on. Agricolan tietosanomat 2/2003.
URL: http://agricola.utu.fi/tietosanomat/numero2-03/tieto.html

Knowledge@Wharton 2003
Knowledge@Wharton: What Is Google Worth? CNET NewsCom. November 29, 2003.
URL: http://news.com.com/2030-1069_3-5112098.html

Lazuly 2003
Pierre Lazuly: Telling Google What to Think. How an Online Search Engine Influences Access To Information. Le Monde Diplomatique. November 2003. (Translated by Gulliver Cragg.)
URL: http://mondediplo.com/2003/11/15google/

Manjoo 2002
Farhad Manjoo: Meet Mr. Anti-Google. Salon.com, Aug. 29, 2002.
URL: http://www.salon.com/tech/feature/2002/08/29/google_watch/

Markoff 2004
John Markoff: The coming search wars. New York Times / CNET News.com. February 2, 2004.
URL: http://news.com.com/2100-1032-5151934.html

Olsen 2002
Stefanie Olsen: The Google Gods. Does search engine's power threaten Web's independence? CNET News.com. October 31, 2002
URL: http://news.com.com/2009-1023-963618.html

Olsen 2003
Stefanie Olsen: Google tests book search. CNET News.com. December 17, 2003.
URL: http://news.com.com/2100-1038_3-5128515.html

Olsen & Hu 2003
Stefanie Olsen & Jim Hu: The changing face of search engines. CNET News.com. March 24, 2003.
URL: http://news.com.com/2100-1032-993677.html

Orlowski 2003
Andrew Orlovski: Google to fix blog noise problem. The Register. May 9, 2003.
URL: http://www.theregister.co.uk/content/6/30621.html

Reuters 2004
[Reuters Limited:] Google continues reign as top brand. CNET

NewsCom. February 3, 2004.
URL: http://news.com.com/2100-1032_3-5152397.html

Rouvari & Ryhänen 2003
Ari Rouvari & Henri Ryhänen: Nelli-portaali - kansallista tiedonhakua. Tietolinja 2/2003.
URL: http://www.lib.helsinki.fi/tietolinja/0203/portaali.html

Sherman & Price 2002
Chris Sherman & Gary Price: The Invisible Web. Uncovering Information Resources Search Engines Can't See. Medford, New Jersey, 2002 (2001).

Sonnenreich 1997
Wes Sonnenreich: A History of Search Engines. 1997.
URL: http://www.wiley.com/legacy/compbooks/sonnenreich/history.html

Tuominen 2003.
Kimmo Tuominen: Tarvitaanko verkkodokumenttien kuvailussa käsityötä. Osa 1: Kontekstuaalinen metadata. Informaatiotutkimus 3/2003.

Agricolan Tietosanomien pääsivulle

Tämän numeron pääsivulle
Lehden arkisto
Lehden toimituskunta
Kaikkien numeroiden sisällysluettelot yhtenä tiedostona

Historian äärelle | Tutkimus, opetus, seurat | Arkistot, kirjastot, museot | Ajankohtaista
Agricolan kartta | Haku Agricolasta | Hakemisto | Uutta!
Tekijät | Palaute | Etusivulle